AM JA999118

日 本 国 特 許 庁
PATENT OFFICE
JAPANESE GOVERNMENT

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日
Date of Application: １９９９年 ７月 ７日

出 願 番 号
Application Number: 平成１１年特許願第１９２６１８号

出 願 人
Applicant (s): インターナショナル・ビジネス・マシーンズ・コーポレイション

１９９９年１１月２６日

特 許 庁 長 官
Commissioner,
Patent Office 近 藤 隆

出証番号 出証特平１１－３０８３３３６

【書類名】　特許願

【整理番号】　JA999118

【あて先】　特許庁長官　殿

【国際特許分類】　G06F 15/00

【発明者】

【住所又は居所】　神奈川県大和市下鶴間１６２３番地１４　日本アイ・ビー・エム株式会社　東京基礎研究所内

【氏名】　那須川　哲哉

【発明者】

【住所又は居所】　神奈川県大和市下鶴間１６２３番地１４　日本アイ・ビー・エム株式会社　東京基礎研究所内

【氏名】　長野　徹

【特許出願人】

【識別番号】　390009531

【住所又は居所】　アメリカ合衆国１０５０４、ニューヨーク州アーモンク（番地なし）

【氏名又は名称】　インターナショナル・ビジネス・マシーンズ・コーポレイション

【代理人】

【識別番号】　100086243

【弁理士】

【氏名又は名称】　坂口　博

【連絡先】　０４６２－７３－３３１８、３３２５、３４５５

【選任した代理人】

【識別番号】　100091568

【弁理士】

【氏名又は名称】　市位　嘉宏

【手数料の表示】

【予納台帳番号】　024154

【納付金額】　　21,000円

【提出物件の目録】

【物件名】　　明細書　1

【物件名】　　図面　1

【物件名】　　要約書　1

【包括委任状番号】　9304391

【包括委任状番号】　9304392

【プルーフの要否】　　要

【書類名】　明細書

【発明の名称】データ分析システム

【特許請求の範囲】

【請求項１】

データから有効な知識を取り出す、データ分析システムであって、

（１）文書データを含むデータから概念を抽出する、概念抽出手段と、

（２）前記抽出された概念から特徴的な概念を抽出する、特徴的概念抽出手段と、

を含む、データ分析システム。

【請求項２】

前記概念抽出手段（１）が、文書データを含むデータからカテゴリ別の概念を抽出する手段であり、前記特徴的概念抽出手段（２）が、前記カテゴリ別の概念において、同一カテゴリに属する概念のうち、対応する別のカテゴリに属する概念の中で占める割合が既定値を超えている概念を抽出する手段である、請求項１記載のシステム。

【請求項３】

前記概念抽出手段（１）が、

データ中の前記文書データを形態素解析する手段と、

前記形態素解析の結果に基づき、前記文書データの文節を生成する手段と、

前記文節内のキーワードの概念を抽出する手段であって、前記文節に対してカテゴリ辞書を適用して、文節内のキーワードに対してカテゴリを付加する手段と、

前記文節を有する文を構文木生成規則に従い構文木を生成する、構文解析手段と、

前記文節内のカテゴリを付加されたキーワードに関し、同一文中でのキーワードの係り受け関係を抽出する、係り受け抽出手段と、

前記カテゴリ別の概念を抽出する手段であって、前記キーワード間の係り受け関係に基づき、係り受けの関係にある各キーワードのカテゴリの組合せを抽出する手段と、

を含む、請求項２記載のシステム。

【請求項４】

前記特徴的な概念を検出する手段（２）が、

ユーザの命令を受け取る入力手段と、前記ユーザの命令を解析する、命令解析手段と、前記解析された命令に従い、前記カテゴリ別の概念を提示し、同一カテゴリに属する概念のうち、対応する別のカテゴリに属する概念の中で占める割合が既定値を超えている概念を他の概念と異なる属性で表示する手段とを有する、請求項２記載のシステム。

【請求項５】

前記特徴的な概念を検出する手段（２）が、さらに、

抽出された概念の相対頻度を計算する手段と、

抽出された概念からキーワードを検索する手段と、

カテゴリ別キーワードの頻度を計算する手段と、

得られた前記相対頻度、検索結果、キーワードの頻度を表示する手段と、

を含む、請求項４記載のシステム。

【請求項６】

データから有効な知識を取り出す、概念抽出方法であって、

データ中の文書データを形態素解析を行う段階と、

前記形態素解析の結果に基づき、前記文書データの文節を生成する段階と、

前記文節に対してカテゴリ辞書を適用して、文節内のキーワードに対してカテゴリを付加する段階と、

前記文節を有する文を構文木生成規則に従い構文木を生成する段階と、

前記文節内のカテゴリを付加されたキーワードに関し、同一文中でのキーワードの係り受け関係を抽出する段階と、

前記キーワード間の係り受け関係に基づき、係り受けの関係にある各キーワードのカテゴリの組合せを抽出する段階と、

を有する、概念抽出方法。

【請求項７】

データから有効な知識を取り出すプログラムを含む媒体であって、該プログラムが、コンピュータに、

データ中の文書データを形態素解析を行う機能と、
前記形態素解析の結果に基づき、前記文書データの文節を生成する機能と、
前記文節に対してカテゴリ辞書を適用して、文節内のキーワードに対してカテゴリを付加する機能と、
前記文節を有する文を構文木生成規則に従い構文木を生成する機能と、
前記文節内のカテゴリを付加されたキーワードに関し、同一文中でのキーワードの係り受け関係を抽出する機能と、
前記キーワード間の係り受け関係に基づき、係り受けの関係にある各キーワードのカテゴリの組合せを抽出する機能と、
を実現させる、プログラムを含む媒体。

【発明の詳細な説明】

【０００１】

【産業上の利用分野】

本発明は、大量のデータから知識を獲得する技術に関し、特にキーワードに付加した同一カテゴリの概念比較により、大量の非定形文からなるデータから注目に値する有効な知識を獲得する方法およびシステムに関する発明である。

【０００２】

【従来の技術】

大量の文書の内容を様々な形で視覚化し分析に役立てようという試みが次第に増えてきている。しかし、従来の方法では、名詞句を中心としたキーワードを抽出し、キーワードの分布を表示（例えば、キーワード間の相関関係を計算し、その結果に基づいて相関の強いキーワードをクラスタ化して表示）する程度の処理しか行っていない。その程度の処理の結果を用いた場合には、ユーザーが様々な観点からデータを絞り込みながら、主観的に注目に値すると思える部分を発見するまで手探りで分析するしかなく、分析過程におけるユーザーの負担が大きい。また、多種多様なキーワードを混合して、まとまりの無いままの状態で扱うため、有効な結果にたどり着くのは難しい。

つまり、テキストをデータマイニングの手法（クラスタリングや相関ルールの分析など）で解析してみようという試みは、データマイニングが注目され始めた

頃から存在するものの、従来の手法では、テキストから抽出した分析単位が、単なる文字列としてのキーワードに過ぎなかった為、有効な結果を得られずに終わってしまうケースが多い。

【０００３】

【発明が解決しようとする課題】

従って、本発明が解決しようとする課題は、大量データから有効な知識を獲得する方法およびシステムを提供することである。

また別の課題は、大量データから注目に値する知識を自動的に見つけ出す方法およびシステムを提供することである。

また別の課題は、大量のデータから有効な知識を獲得するにあたり、ユーザビリティに優れた分析方法およびシステムを提供することである。

【０００４】

【課題を解決するための手段】

上記課題を解決するために、本発明は文書を含む大量のデータから特異な特徴を有する概念を抽出することにより、有効な知識を獲得する方法およびシステムを提供する。

本発明は、概念抽出装置と特徴的概念抽出装置と有する。前記概念抽出は文書データを含むデータからカテゴリ別の概念を抽出する。前記特徴的概念抽出装置は前記抽出した概念の中から特徴的な概念を抽出する装置であって、前記カテゴリ別の概念において、同一カテゴリに属する概念のうち、対応する別のカテゴリに属する概念の中で占める割合が既定値を超えている概念を抽出する。

より、詳細には、概念抽出装置では、語彙辞書や文法知識を利用した形態素解析やカテゴリ辞書を用いた曖昧性解消手法を用いて、非定型テキストから、カテゴリ別の概念を抽出する。特徴的概念抽出装置では、カテゴリとカテゴリの各組み合わせにおいて、同じカテゴリの概念のうち、対応する別のカテゴリの概念の占める割合の比や差が既定値（しきい値）を超えているものを見つけ出す。さらにカテゴリとカテゴリの各組み合わせを表形式で表示し、該表中において注目すべき項目を他と異なる属性で表示したり、リストアップして提示するなどの処理を行う。

【０００５】

【発明の実施の形態】

図１は、本発明のデータ分析システム概要を示すブロック図である。なお実施例としてコンピュータ製品に関する電話の問い合わせデータに基づき、これを解析し、特徴ある概念を抽出するデータ分析システムについて説明する。図１のブロック１１０は問い合わせデータ１５０を入力としてラベル付きデータ１６０を出力するデータ変換部である。問合せデータ１５０に含まれる非定型のデータと、同じく含まれる定型のデータを同一に保持する形式のデータを作成する。ブロック１２０は、入力された文に対してカテゴリ辞書１７０を用いてカテゴリを付加し、カテゴリを付加されたキーワードのうち、同じ文中で係り受け関係のあるものを、より具体的な意味を表現する概念（ラベル・概念付きデータ１８０）として抽出する、概念抽出部である。ブロック１３０は、ラベル・概念付きデータ１８０を入力として、特徴のある概念を検索・抽出する検索・特徴検出部１３０である。このブロックは検索・特徴抽出を効率良く行えるようにするため、データ全体もしくは部分集合における頻度分布などの統計情報を作成・保持する。同じカテゴリに属する概念は同じような振る舞いをする（同じような出現傾向を持ち、同じような概念と共起する）傾向があると想定し、各概念の振る舞いを全体、もしくは部分集合の値と比較することで、効率良く、注目すべき情報を検索・検出する。さらにこの検索・特徴検出部１３０は、統計情報を視覚的に表示する機能を有し、特徴ある概念の分布差異の提示を行う。

【０００６】

上記各ブロック１１０、１２０、１３０について以下に詳細に説明する。

［データ変換部１１０］

まずブロック１１０に入力される問い合わせデータの例は以下のようなものである。

--------- 問い合わせデータ例 ----------

1999/01/01

0000001

タイトル：ノートで日本語が使えない

マシンタイプ：製品Ａ

問題種別：総合案内

ＣＡＬＬ種別：案内

回答・対応種別：窓口対応

解決期間：１日

通話時間：２１分

ご質問：ノートパソコンで日本語が使えないので、ＯＳを再インストールしたのですが、それ以降、モデムとイーサネットカードが使えなくなってしまいました。

--------- 問い合わせデータ例 --終わり--------

このように、問合せデータは日付や定型項目からの選択肢、また連続値や離散値、自由に記入できるタイトルや文章など、さまざまな形式の項目から構成されたデータからなる。

【０００７】

ブロック１１０は上記のような問い合わせデータ１５０を入力とし、下記のようなラベル付きデータに変換する。

--------- ラベル付きデータ ----------

ID199901010000001

TITＰで日本語が使えない

KWM1MＴ：製品Ａ

KWQ3ＴＣ：総合案内

KWQ4ＴＤ：案内

KWQ2ＰＴ：窓口対応

KWP3ＳＤ：１日

KWP4ＣＭ：２１分

CTQ：ノートパソコンで日本語が使えないので、ＯＳを再インストールしたのですが、それ以降、モデムとイーサネットカードが使えなくなってしまいました。

--------- ラベル付きデータ --終わり---

このように、非定形データを含む問い合わせデータを上記のようなラベル付きデ

ータに変換することにより、さまざまな型のデータを同じ形式に変換する。上の例ではTIはタイトル、ＭＴはマシンタイプ、ＴＣは問題種別、ＴＤはＣＡＬＬ種別、ＰＴは回答・対応種別、ＳＤは解決期間、ＣＭは通話時間、ＣＴは元の問合せ内容、KW+2byteは項目の種類を示す。図２にデータ変換部のフローチャートを示す。ステップ２１０で問い合わせデータ１５０を読み込み、ステップ２２０でデータの終了であるかを判断し、そうでなければステップ２３０で形式を変換する。データが終了した場合はステップ２４０で変換を終了する。

【０００８】

［概念抽出部１２０］

図３にブロック１２０の概念抽出部を説明する。ブロック３１０は、ラベル付きデータ１６０内の文に対して、形態素解析を行う、形態素解析装置である。次にブロック３２０は、ブロック３１０で形態素に分解された文について形態素を文（あるいはは特定の文脈）に出現する順に文節生成規則に従って文節を決定する、文節生成装置である。形態素が付属語である場合や明らかに文節が切れると判断されたところで文節を区切りこれを文末になるまで行う。

【０００９】

次にブロック３３０は、ブロック３２０で生成された文節に対してカテゴリ辞書３４０を適用し文節内のキーワードに対してカテゴリを付加する辞書適用装置である。キーワードにカテゴリを付加することで単なる文字列ではなく意味を持った概念として扱う。（例えば、「ワシントン」という文字列を単なるキーワードとして扱えば、人名も地名も区別できず、有効な分析が行えないが、[人名] [地名]といったカテゴリを付加することで、意味を持つことが出来る）体言（名詞類の語）は（元表現 品詞 概念 カテゴリ）のデータで表現されるカテゴリ辞書を参照してカテゴリを付加する。述語に関しては体言と同様にカテゴリ辞書を用いると共に、付属語の情報から[問題] [要望] [疑問]といったカテゴリ付けを行う。例えば、「壊れた」という動詞は（壊れる [動詞] 故障 [問題]）というデータがカテゴリ辞書にあれば、[問題]というカテゴリに属する「故障」という概念として抽出されるが、「…できない」「…したい」という表現は、カテゴリ辞書を参照せずに[問題] [要望]として解釈することができる。上記カテゴリ辞書１７

０の構造の例を以下に示す。

--------- カテゴリ辞書の例 ----------

ノートパソコン	固有名詞	ノートパソコン
N1		
ＯＳ	固有名詞	オペレーティングシステム
N2		

--------- カテゴリ辞書の例 -終わり---

上記辞書においてカテゴリはN1（ハードウェア）、N2（ソフトウェア）というように対応付けられている。

【００１０】

ブロック３５０はブロック３３０でキーワードに対してカテゴリが付加された文節を有する文を簡単な構文木生成規則で構文木を生成する構文木解析装置である。

【００１１】

ブロック３６０は文節内のカテゴリを付加されたキーワードのうち同じ文中で係り受け関係のあるものを、より具体的な意味を表現する概念として抽出する係り受け抽出装置である。このブロック３６０では構文解析装置３５０の構文解析結果により得られるキーワード間の係り受け関係に基づき、係り受けの関係にある各キーワードのカテゴリの組合せを概念（ラベル・概念付きデータ３７０）として抽出する。ラベル・概念付きデータ３７０の例を以下に示す。

--------- ラベル・概念付きデータの例 ----------

ID199901010000001

TIノートで日本語が使えない

KWM1ＭＴ：製品Ａ

KWQ3ＴＣ：総合案内

KWQ4ＴＤ：案内

KWQ2ＰＴ：窓口対応

KWP3ＳＤ：１日

KWP4ＣＭ：２１分

CTQ：ノートパソコンで日本語が使えないので、ＯＳを再インストールしたのですが、それ以降、モデムとイーサネットカードが使えなくなってしまいました。
KWN1ノートパソコン
KWN0日本語
KWV2使えない
KWW6ノートパソコン…使えない
KWN2ＯＳ
KWV6再インストールする
KWWDＯＳ…再インストールする
KWN1モデム
KWN1イーサネットカード
KWV2使えない
KWW6モデム…使えない
KWW6イーサネットカード…使えない
--------- ラベル・概念付きデータの例 --終わり---
上記のようにラベル・概念付きデータ３７０は、データ変換部で得られたラベル付きデータ１６０に、概念抽出部１２０で抽出されたデータを追加した形を取り、ラベル付きデータ１６０と同じく同一形式のデータとなる。

【００１２】

図１２に本発明の概念抽出部の流れを実際の文に基づき説明する。まずステップ１２１０において、入力文”MODEMとイーサネットカードが使えない。”が入力された場合、ステップ１２２０において、形態素解析装置３１０により、文を単語に区切り、品詞番号が付加される。例えば、
104…固有名詞、 81…格助詞「と」、 75…格助詞「が」、 10…動詞語幹、 44…???、 51…打ち消しの助動詞「ない」、100…句読点、などのような対応表を用いて形態素解析を行う。これにより入力文”MODEMとイーサネットカードが使えない。”は、以下のように変換される。
[MODEM，104] [と，81] [イーサネット，104] [カード，104] [が，75] [使，10] [え，44] [ない，51] [。，100]

【００１３】

次にステップ１２３０、およびステップ１２４０で文節生成が行われる。ステップ１２３０の文節生成１では、形態素解析された文の単語列を文節にまとめる操作を行う。あらかじめ”{81，75，100，…}で文節を区切る”というルールを用意しておき、このルールを文頭から適用して、文頭から文節ごとに区切っていく操作を行う。上記入力文の場合、先頭から3文節あるが、各文節の先頭の単語がそれぞれ、名詞・名詞・動詞であることから、それぞれ順に、体言句・体言句・用言句であると判断される。その結果、入力文は以下のように変換される。

{[MODEM，104] [と，81]}

{[イーサネット，104] [カード，104] [が，75]}

{[使，10] [え，44] [ない，51] [。，100]}

【００１４】

次にステップ１２４０の文節生成２では、文節生成１で区切られた文節に対して、すべての文節を自立語と付属語の組にする働きを行う。体言句は、名詞が複数含まれる場合は先頭から名詞を結合するようにする。例えば、{[イーサネット，104] [カード，104]→[イーサネットカード，104]）のようにする。その後自立語の品詞コードを、一般名詞句を示すN1に書き換える。用言句は、付属語列（[え，44] [ない，51] [。，100]）を解析し、否定の情報を示す[ない，51]を取り出し、動詞の語幹[使，10]は[え，44]と結合して、終止形「使える」にする。品詞コードを一般動詞を示すV1にし、否定の情報を付加して-V1とする。その結果、入力文は以下のように変換される。

{[MODEM，N1] [と，81]}

{[イーサネットカード，N1] [が，75]}

{[使える，-V1] [。，100]}

次にステップ１２５０で、カテゴリ辞書を用いて、自立語と付属語の組に分解された文節に対してカテゴリ付を行う。ここで適用される辞書は、以下の3つである。

（MODEM　　N1　　モデム　　NA）

（イーサネットカード　　N1　　イーサネットカード　　NA）

（使える　　-V1　　使えない　　VC）

なおNAはハードウェアを意味し、VCは問題を意味する。 その結果、入力文は以下のように変換される。

{[モデム，NA] [と，81]}

{[イーサネットカード，NA] [が，75]}

{[使えない，VC] [。，100]}

【００１５】

次にステップ１２６０で、カテゴリ付された文節に文節からなる文に基づき構文木を生成する。この時の係り受けルールの形式は（係り受け元文節の自立語，係り受け元文節の付属語 ，係り受け先文節の自立語 ，係り受け先文節の付属語）となっている。このルールを文頭の文節1 {[モデム，NA] [と，81]} から適用する。文節n番目に対してn+1番目から最終文節Nまで適用する(n = 1 ～ N-1)。係り受けルール中に(NA，81，VC，*)というルールがあるので、{[モデム，NA] [と，81]}と{[使えない，VC] [。，100]}に係り受けの関係があると判断される。なおルール中の*はすべての品詞またはカテゴリーにマッチするという意味である。これを(n = 1 ～ N-1)で行い、係り受け情報を含む文節を有効グラフとして表し、(係り受け元，係り受け先，自立語，カテゴリー，付属語の品詞番号)という形式に変換する。その結果、入力文は以下のように変換される。

(1，3，"モデム"，NA，81)

(2，3，"イーサネットカード"，NA，75)

(3，NULL，"使えない"，VC，100)

【００１６】

最後に、ステップ１２７０において、構文解析された文を入力として、係り受け抽出抽出ルールに従って係り受けを抽出する。抽出ルールは任意の長さのカテゴリの列からなる。例えば(カテゴリ１，カテゴリ２，... ，カテゴリn) という形式である。文節番号1からNまでの係り受けを見て、係り受け抽出ルールに(NA，VC)というルール(n = 2)があるので、「モデム…使えない」と「イーサネットカード…使えない」という2つの係り受けが抽出される。

結局、元の文書である「モデムとイーサネットカードが使えない」という文章か

ら

「モデム …使えない」［ハードウェア…問題］

「イーサネットカード …使えない」［ハードウェア…問題］

という概念情報が取り出されたことになる。このようにして取り出された概念情報は、ラベル・概念付きデータベース１８０に登録される。

【００１７】

図４に、本発明の概念抽出部１２０の処理のフローチャートを示す。

ステップ４２０でラベル付きデータ１６０内の文章Ｔを形態素W0～Wmに分割する。ここで形態素Wは、文字列wと品詞pで表される。すなわちW = ｛ w，p ｝である。（以上が形態解析装置の処理である）

【００１８】

次にステップ４３０で全単語が文節に変換されたかどうかを判断し、そうであれば処理はステップ４４０へ移り、そうでなければステップ４３２で単語Ｗｎが付属語または句読点がどうかを判断し、その結果がＮｏであればステップ４３４で、文節Ｐｉに単語Ｗｎを追加する。ここで文節Pは、1つ以上の連続した単語Wの集合で P = ｛ W ｝ = ｛ ｛ w，p ｝ ｝である。そして処理はステップ４３０に戻る。ステップ４３２の判断の結果がＹｅｓであればステップ４３６で文節Ｐｉに単語Ｗｎを追加した後、その次の文節を用意する。処理はその後ステップ４４０に移る。ステップ４４０で、全文節に対する処理が必要かどうかを判断し、必要であれば処理はステップ４５０へ移り、費用でなければステップ４４２において、文節をP = ｛ ｛ w，p ｝ ｝の形式から P’ = ｛ {w1，p1} ｛ w2，p2} ｝に変換する。（ここでw1は自立語w2は付属語である）例えば P = ｛ ［国際，名詞］［情勢，名詞］［は，助詞］｝であれば、名詞句はひとつにまとめ、 P’ = ｛ ［国際情勢，名詞］［は，助詞］｝ とする。そして処理はステップ４４０へ戻る。（以上が文節生成装置３２０の処理である）

【００１９】

ステップ４５０では、自立語に対して辞書引きが終了したかどうかを判定する。その判定結果がNoであれば、ステップ４５２で、((w1，p1) == (wa， pa))のエントリに対し (w1 = wb，p1 = pb) とする。同時に用言の処理を行う。この時

用いられる辞書４５４は[wa pa wb pb]のエントリの集合からなり、それぞれ[元表現，品詞，概念，カテゴリ]を表す。ここでの概念というエントリは、置き換え表現を意味する。元表現が「ＰＣ」であれば置き換え表現は「パソコン」となる。また全エントリはハッシュ構造で格納されているので高速にアクセスできる。例えばエントリ集合は、[マシン，名詞，機械，ハードウェア]、[壊れる，動詞，壊れる，問題]等である。この辞書を用いて、名詞句 P' = { [マシン，名詞] [が，助詞] }であれば、 P'' = { [機械，ハードウェア] [が，助詞] } となる。用言の句 P' = { [壊れる，動詞] [ない，助詞] }であれば、動詞→問題となるはずだが、「ない」があるので、P'' = { [壊れない，好評] [NULL，NULL] }とする 。また辞書に該当するエントリがなければ何もしない。ステップ４５０の判定結果がYesであれば処理はステップ４６０へ進む。（以上が辞書適用装置３３０の処理である）

【００２０】

ステップ４６０で、構文木が完成したかどうかが判定される。構文木が完成した場合、処理はステップ４７０へ進む。構文木が完成していない場合、処理はステップ４６２で、一般的に行われる構文解析を行い、結果として、PnとPkがリンクされる。（以上が構文解析装置３５０の処理である。）

【００２１】

ステップ４７０で、係り受けの抽出が終了したかどうかを判断する。もし終了していなければ処理はステップ４７２へ進み、ルールに基づき、Pnに対してリンクされた２項関係を抽出し、ラベル・概念付きデータベース１８０に登録する。この時、係り受けルール４７４を参照する。係り受けルール４７４のルールは[px py]のエントリの集合からなり、それぞれ[係り受け元カテゴリ，係り受け先カテゴリ]を表す。例えば、Pn = { [機械，ハードウェア] [が，助詞] } ，Pk = { [壊れる，問題] [NULL，NULL] } （nとkには係り受け関係があるものとする）であれば、上のルールを用いて、[ハードウェア，問題] → [機械，壊れる] が抽出され、ラベル・概念付きデータベース１８０に登録される。ステップ４７０の判断がYesであれば処理はステップ４８０で終了する。（以上が係り受け抽出装置３６０の処理である。）

【００２２】

［検索・特徴検出部１３０］

図５に検索・特徴抽出部１３０のブロック図を示す。検索・特徴抽出部１３０は入力（命令）装置５７０、表示部５１０、キーワード検索装置５４０、命令解析装置５２０、カテゴリー別キーワード頻度計算装置５５０、相対頻度計算装置５３０の各ブロックから構成される。なおキーワード検索装置５４０及びカテゴリー別キーワード頻度計算装置５５０は、ラベル・概念付きデータベース１８０にアクセスして概念情報の検索を行う。好適にはラベル・概念付きデータベース１８０は、ラベル・概念付きデータに対してインデックスを生成しておき、高速な検索を可能にしておく。

命令解析装置５２０は、入力（命令）装置５７０から受け取る命令を解析して、各装置にキーワード・パラメータを送る。入力（命令）装置５７０は、図１４におけるキーボード６、マウス７などに相当し、ユーザの指示に従い、データ分析システムに対して所望の検索、表示を行わせるために使用される。相対頻度計算装置５３０は、文書の全体、または部分集合に対して相対頻度を計算する装置である。ここで相対頻度とは、全体または任意の集合Xに対して含まれる各キーワードと、任意の集合Yに含まれるキーワードの集合を比較することにより計算される。

【００２３】

キーワード検索装置５４０は、命令解析装置５２０の出力であるキーワードまたはキーワードの組を入力として、文書の全体、または部分集合に含まれるキーワードの数とキーワードを含む文書のＩＤを得る装置である。該装置によりキーワードを含む文書集合を絞り込むことが出来る。

【００２４】

カテゴリー別キーワード頻度計算装置５５０は、命令解析装置５２０の出力に従い、文書全体、または部分集合に含まれるキーワードの数をカテゴリー別に、頻度順に得る装置である。以下に該装置の出力例を示す。（下の例でINPUTではカテゴリを指定、N1は[ハードウェア]を表すカテゴリであり、OUTPUTでは[キーワード 出現頻度]となっている。

--------- カテゴリー別キーワード頻度計算装置の出力例 ----------

[INPUT]CATEGORY	N1
[OUTPUT]ハードディスク	2033
[OUTPUT]モニタ	1432
[OUTPUT]プリンタ	1001
[OUTPUT]モデム	420
[OUTPUT]スキャナ	212
[OUTPUT]イーサネットカード	143
[OUTPUT]マウス	3

--------- カテゴリー別キーワード頻度計算装置の出力例 --終わり--

【００２５】

表示部５１０は図６に示す表示領域（１）と図７に示す表示領域（２）を含むＧＵＩ画面から構成される。ユーザが表示部５１０中に表示された種々の項目を入力（命令）装置５７０により適宜選択したり、検索のためのパラメータなどを入力することにより、種々の結果（頻度表示、検索結果表示など）を表示部５１０に表示する。例えば図６では、概念Ａを横軸、概念Ｂを縦軸にとった2次元の表である。Ａの列に対して特徴的なＢを表すセルが他のセルとは異なる属性で表示されている（例えば概念A3の中で特徴的な概念はB1である。また複数の概念が特徴的なこともある）。上記異なる属性で表示されたセルをクリックすることで、概念Axと概念Byを含む集合の検索が行われ、そこで得られた集合に対しても再び検索を行うことができる。またはこの特徴概念の表示を行うことができる。表示部５１０の図７に示す表示領域（２）では、概念[ソフトウェア]に含まれるキーワードのリストを表示してある。１つのキーワードに対して２つのグラフがあるが、上は頻度、下は相対頻度を表している。さらに頻度・相対頻度のどちらかで並び替えることができる。これらの表示部５１０における特徴的な概念の抽出及び表示の流れを以下により詳細に説明する。

【００２６】

図８に表示領域（１）における、特徴的な概念の抽出及び表示のフローチャートを示す。ステップ８２０で概念Ａと概念Ｂを選択する。ここで概念A，Bはそれぞ

れ表示領域（１）でのx，y軸になる。後の計算で、概念Aの要素Axに関して特徴的なものを表示するので、比較したい概念をBにセットする。ステップ８４０で概念Aを装置５４０に入力し、概念A中に含まれるキーワードが頻度順に取り出される。また概念Bを装置５４０に入力し、概念B中に含まれるキーワードが頻度順に取り出される。次にステップ８５０で、AとBの組合せが終了したかどうかが判断される。終了していなければステップ８５５で、（Ai & Bj）を装置５４０に入力し、結果をPijに保存する。次にステップ８６０で、正規化が終了したかどうかを判断する。正規化が終了していなければステップ８６５において、装置５３０でBNiを1として正規化する。すなわち（P'ij = Pij / BNi)となる。図６における表示例中の%表示がそれにあたる。ただし、１つの文章に複数の種類のキーワードが表れるので、これらの値を全部足しても1にはならない。次にステップ８７０で、相対頻度の計算が終了したかどうかを判断する。計算が終了していなければステップ８７５で、P'ijをy軸方向に比較して相対頻度を装置５３０で調べる。好適にはP'ij(０ ＜ ｉ ＜ a)の分布の分散を調べ、各P'ijが分散の何倍になっているかを相対頻度とする。次に処理はステップ８８０へ進み、相対頻度の値によって、表示領域（１）の表示色を変え、2次元に表示する。次にステップ８９０で、目立っている概念A中のキーワードfaと概念B中のキーワードfbの交点をクリックする。最後にステップ８９８で装置５４０により、概念Aのfaと概念Bのfbでデータ集合が絞り込まれる。

【００２７】

図９に表示領域（２）における、本発明の別の特徴的な概念の抽出及び表示のフローチャートを示す。まずステップ９１５で表示領域（２）の左側で概念Aを選択する。次にステップ９２０で装置５５０により、概念A中に含まれるキーワードが頻度順に取り出される。そしてステップ９３０で正規化が終了したかどうか判断される。終了していなければステップ９４０で相対頻度を計算する。ここでの相対頻度の計算は、あるキーワードが全体または任意の部分集合Xの中に含まれる割合と、任意の部分集合Yに含まれる割合 Y/Xで求める。この値が大きいほど、部分集合Yに特徴的なキーワードと考える。ステップ９３０において正規化が終了した場合ステップ９５０で、表示領域（２）の右側に頻度と相対頻度を同

時に表示し、頻度または相対頻度で並び替えて表示する。次にステップ９６０で表示領域（２）の右側でキーワードfaを選択する。そして最後にステップ９７０で入力装置５７０を用いて選択（クリック）することで、装置５４０による絞込みが行える。

【００２８】

上記２つの検索・抽出方法を組み合わせることにより、特徴的な概念を効果的に見つけることが可能となる。例えば表示領域（２）において、概念［月］について検索し、「11月」を選択する。次に表示領域（１）において、［製品名（コンピュータの機種名）］を縦軸（比較したい対象）、［問題］を横軸にとる（図１０参照）。すると「遅い」という［問題］に関して、特徴的な［製品名］が２つマークされている。そして、より相対頻度の高い「製品A」に注目し、「製品A」と「遅い」の交わっているところを選択（クリック）する。（図１０参照）絞り込まれた状態で表示領域（２）において、概念［ハードウェア］について見てみる。すると、頻度の上位２番目には「ハードディスク」があり、相対頻度も高く（7.18倍）、この製品Ａに関しては「ハードディスク」が特有の問題を持っていることを推測できる。（図１１参照）

【００２９】

図１３に本発明で最も特徴的なＧＵＩの例を示す。図１３において表示領域（２）の左側で概念[月]を選ぶ。すると装置５５０により、概念[月]中に含まれるキーワードが頻度順に取り出される。そして装置５３０により得られたキーワードの相対頻度が計算され、表示領域（２）の右側に[月]に含まれる概念が表示される。次に、表示領域（２）の右側でキーワード「11月」を選択する。これにより装置５４０により、概念[月]の「11月」でデータ集合が絞り込まれる。

【００３０】

表示領域（１）で概念[問題]X軸に設定し、Y軸に[機種名]をとる。すると装置５５０で概念[問題]中に含まれるキーワードが頻度順に取り出される。また装置５５０で概念[機種名]中に含まれるキーワードが頻度順に取り出される。これらは検索・正規化され2次元に表示される。そして目立っている概念[問題]の「遅い」と概念[機種名]の「製品A」の交点をクリックする。すると装置５４０により

、概念[問題]の「遅い」と概念[機種名]の「製品A」でデータ集合が絞り込まれる。

【００３１】

表示領域（２）の左側で概念[ハードウェア]を選ぶ。装置５５０により、概念[ハードウェア]中に含まれるキーワードが頻度順に取り出される。そして装置５３０により得られたキーワードの相対頻度が計算される。表示領域（２）の右側に[ハードウェア]に含まれる概念が表示される。結局、上位2番目には「ハードディスク」があり、相対頻度も高く、この製品に関しては、「ハードディスク」が特有の問題ではないかと考えられる。なお上位１番目は、本製品の製品番号などが該当するのでこれらは容易に無視できる。

【００３２】

図１４に本発明において使用されるデータ分析システムのハードウェア構成例を示す。システム１００は、中央処理装置（ＣＰＵ）１とメモリ４とを含んでいる。ＣＰＵ１とメモリ４は、バス２を介して、補助記憶装置としてのハードディスク装置１３（またはＣＤ－ＲＯＭ２６、ＤＶＤ３２等の記憶媒体駆動装置）とＩＤＥコントローラ２５を介して接続してある。同様にＣＰＵ１とメモリ４は、バス２を介して、補助記憶装置としてのハードディスク装置３０（またはＭＯ２８、ＣＤ－ＲＯＭ２９、ＤＶＤ３１等の記憶媒体駆動装置）とＳＣＳＩコントローラ２７を介して接続してある。フロッピーディスク装置２０はフロッピーディスクコントローラ１９を介してバス２へ接続されている。好適にはラベル付きデータ１６０、辞書３４０、ラベル・概念付きデータ３７０はこれら補助記憶装置に記憶される。

【００３３】

フロッピーディスク装置２０には、フロッピーディスクが挿入され、このフロッピーディスク等やハードディスク装置１３（またはＣＤ－ＲＯＭ２６、ＤＶＤ３２等の記憶媒体）、ＲＯＭ１４には、オペレーティングシステムと協働してＣＰＵ等に命令を与え、本発明を実施するためのコンピュータプログラム、オペレーティングシステムのコード若しくはデータを記録することができ、メモリ４にロードされることによって実行される。これらコンピュータ・プログラムのコー

ドは圧縮し、または、複数に分割して、複数の記録媒体に記録することもできる。

【００３４】

システム１００は更に、ユーザ・インターフェース・ハードウェアを備え、入力をするためのポインティング・デバイス（マウス、ジョイスティック等）７またはキーボード６や、ディスプレイ１２を有することができる。好適にはポインティング・デバイス７を用いて、ディスプレイ１２に表示された表示領域（１）、表示領域（２）の項目の選択、変更や、パラメータの入力をGUIで行う。また、パラレルポート１６を介してプリンタを接続することや、シリアルポート１５を介してモデムを接続することが可能である。このシステム１００は、シリアルポート１５およびモデムまたは通信アダプタ１８(イーサネットやトークンリング・カード)等を介してネットワークに接続し、他のコンピュータ、サーバ等と通信を行う。本発明にデータ分析システムは、必要なデータベースを、通信回線を介して接続された外部のサーバ、WAN、LAN内にあるローカルなサーバなどに記憶してもよい。どちらにしても本発明の実施の制限となるものではない。またシリアルポート１５若しくはパラレルポート１６に、遠隔送受信機器を接続して、赤外線若しくは電波によりデータの送受信を行ってもよい。

【００３５】

スピーカ２３は、オーディオ・コントローラ２１によってＤ／Ａ（デジタル／アナログ変換）変換されたサウンド、音声信号を、アンプ２２を介して受領し、サウンド、音声として出力する。また、オーディオ・コントローラ２１は、マイクロフォン２４から受領した音声情報をＡ／Ｄ（アナログ／デジタル）変換し、システム外部の音声情報をシステムにとり込むことを可能にしている。ViaVoice（IBM商標）などのアプリケーションを用いて、本発明のＧＵＩコマンド部の操作を音声コマンドによる操作で代用してもよい。さらにホームページ・リーダー(IBM商標）などアプリケーションを用いて、表示された特徴的概念を有する検索結果などを音声で読み上げるようにしてもよい。

【００３６】

このように、本発明のデータ分析システムは、通常のパーソナルコンピュータ

（ＰＣ）やワークステーション、ノートブックＰＣ、パームトップＰＣ、ネットワークコンピュータ、コンピュータを内蔵したテレビ等の各種家電製品、通信機能を有するゲーム機、電話、ＦＡＸ、携帯電話、ＰＨＳ、電子手帳、等を含む通信機能有する通信端末、または、これらの組合せによって実施可能であることを容易に理解できるであろう。ただし、これらの構成要素は例示であり、その全ての構成要素が本発明の必須の構成要素となるわけではない。

【００３７】

【発明の効果】

本発明により、大量データから有効な知識を獲得する方法およびシステムが提供される。また有効な知識を獲得するための、ユーザビリティに優れたＧＵＩによる分析方法およびシステムが提供される。

【図面の簡単な説明】

【図１】

本発明のデータ分析システム概要を示すブロック図である

【図２】

データ変換部のフローチャートである。

【図３】

概念抽出部のブロック図である。

【図４】

概念抽出部のフローチャートである。

【図５】

検索・特徴抽出部のブロック図である。

【図６】

表示部における表示領域（１）の例である。

【図７】

表示部における表示領域（２）の例である。

【図８】

表示領域（１）における概念の抽出及び表示のフローチャートである。

【図９】

表示領域（２）における概念の抽出及び表示のフローチャートである。

【図１０】

表示領域（１）の典型的な表示例である。

【図１１】

表示領域（２）の典型的な表示例である。

【図１２】

概念抽出部の処理を具体的な文を用いて説明した図である

【図１３】

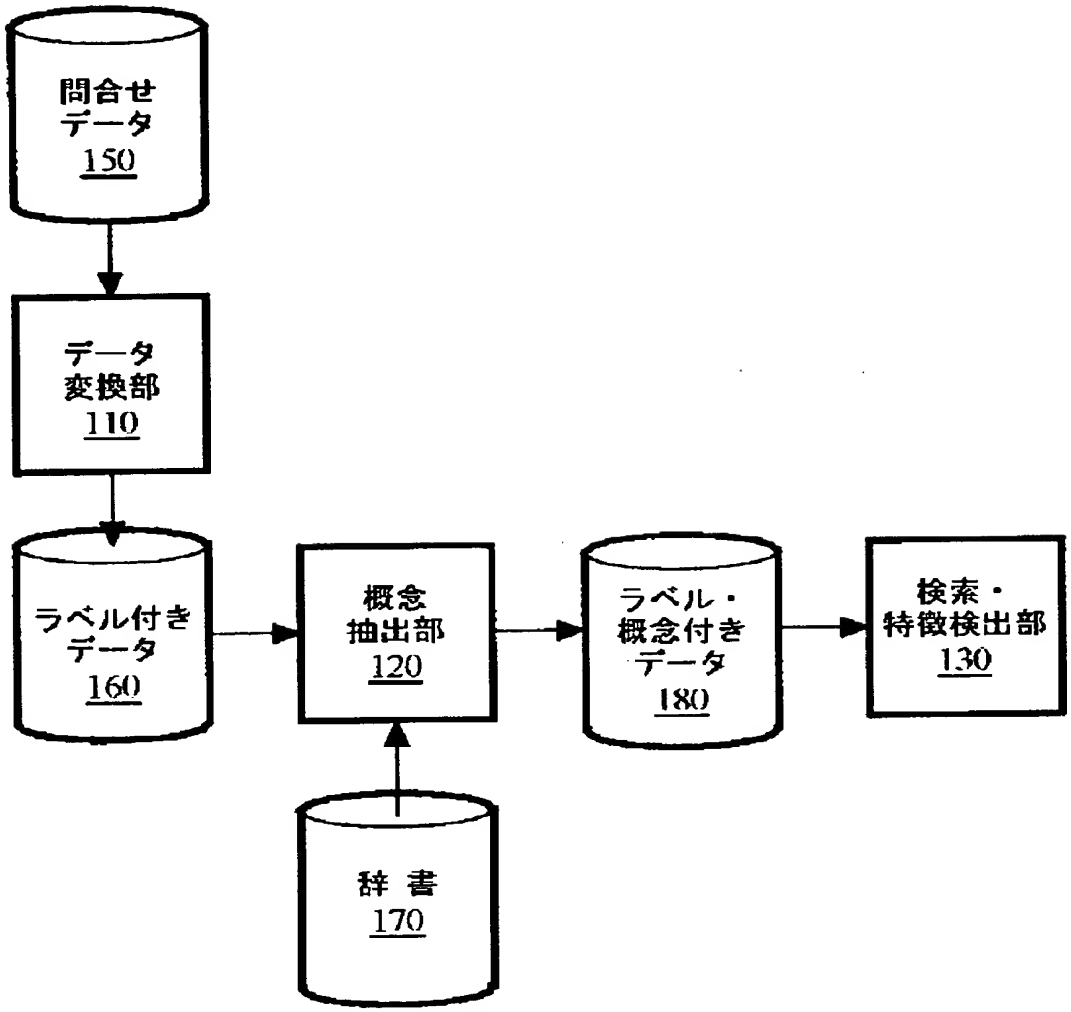
表示領域（１）と表示領域（２）を含むＧＵＩ画面の表示例である。

【図１４】

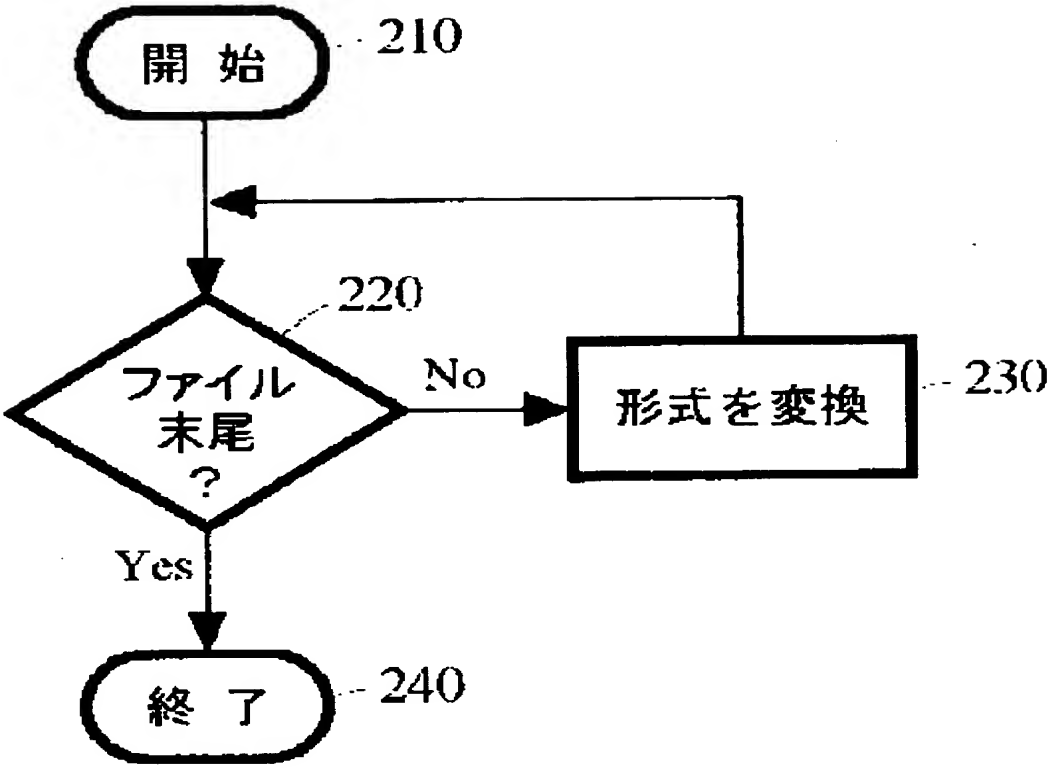
本発明で用いるハードウェア構成の実施例である。

【書類名】　　図面

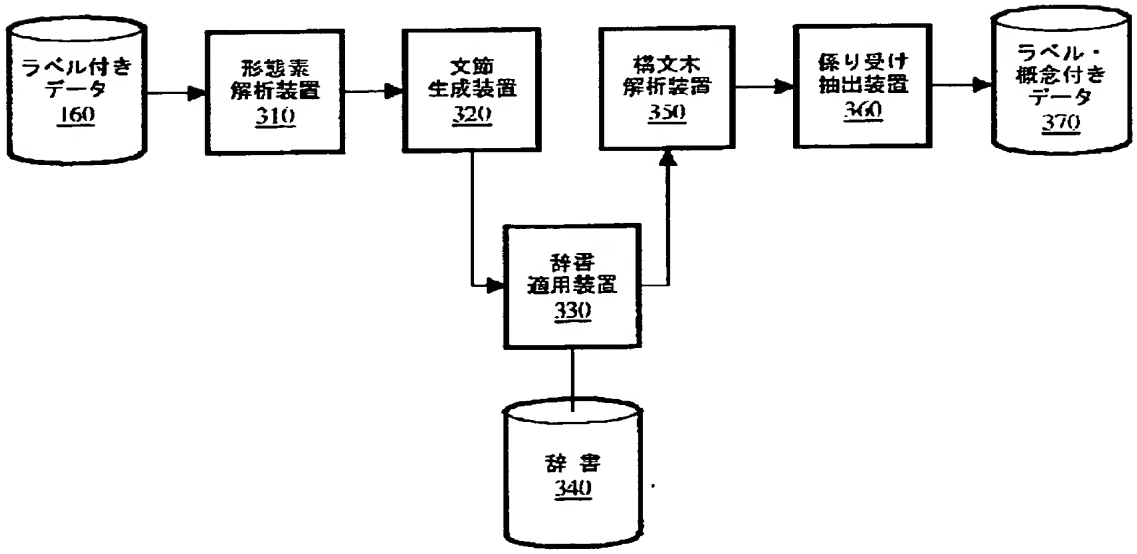
【図１】


問合せ
データ
150
データ
変換部
110
ラベル付き
データ
160
概念
抽出部
120
辞 書
170
ラベル・
概念付き
データ
180
検索・
特徴検出部
130

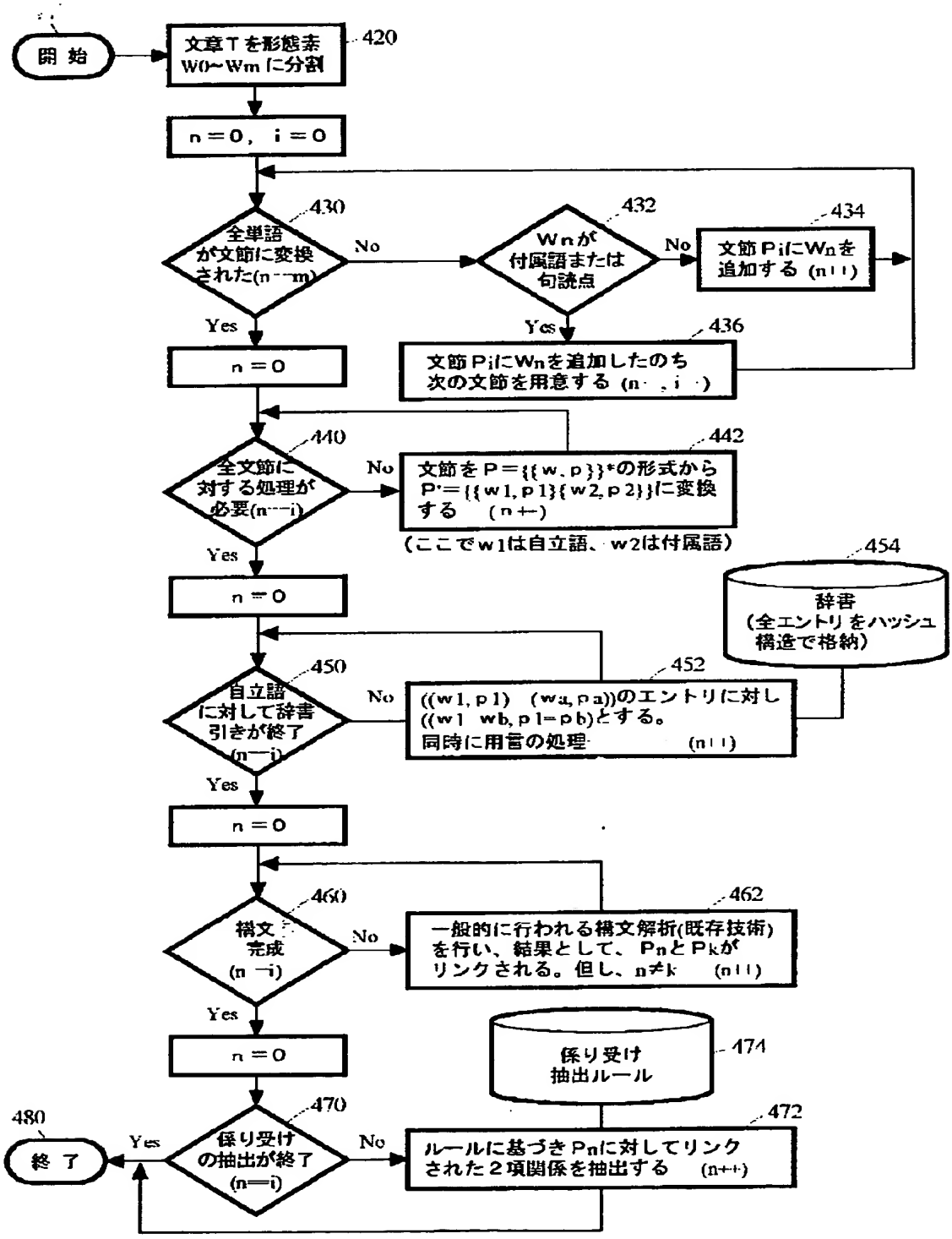
【図２】


開 始
210
220
ファイル
末尾
？
No
形式を変換
230
Yes
終 了
240

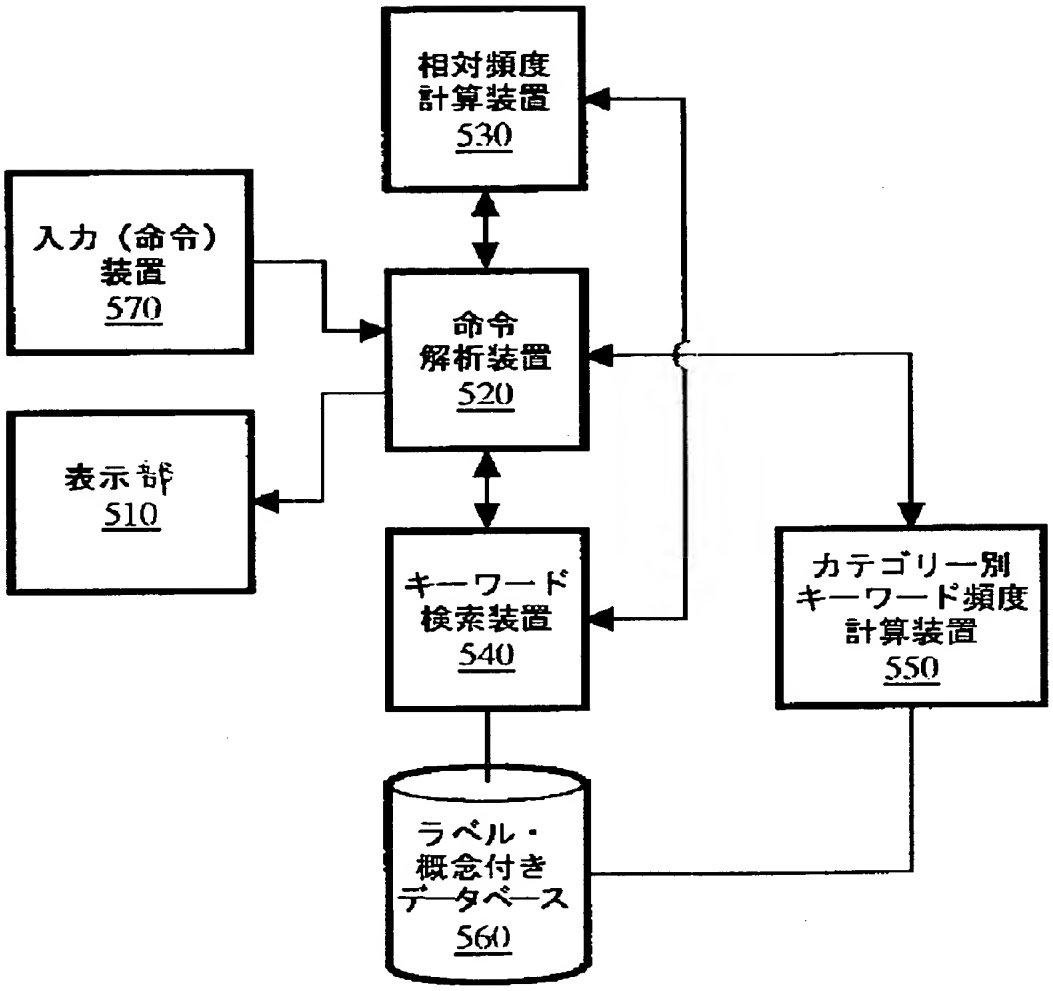
【図３】


ラベル付き
データ
160
形態素
解析装置
310
文節
生成装置
320
構文木
解析装置
350
係り受け
抽出装置
360
ラベル・
概念付き
データ
370
辞書
適用装置
330
辞 書
340

【図４】


開 始
文章Tを形態素 W0~Wmに分割
420
n＝0， i＝0
430
全単語が文節に変換された(n=m)
No
Yes
n＝0
432
Wnが付属語または句読点
No
Yes
434
文節PiにWnを追加する (n++)
436
文節PiにWnを追加したのち次の文節を用意する (n++, i++)
440
全文節に対する処理が必要(n=i)
No
Yes
n＝0
442
文節をP＝{{w, p}}*の形式からP'＝{{w1, p1}{w2, p2}}に変換する (n++)
（ここでw1は自立語、w2は付属語）
454
辞書（全エントリをハッシュ構造で格納）
450
自立語に対して辞書引きが終了(n=i)
No
Yes
n＝0
452
((w1, p1) (wa, pa))のエントリに対し((w1 wb, p1=pb)とする。同時に用言の処理 (n++)
460
構文完成 (n=i)
No
Yes
n＝0
462
一般的に行われる構文解析(既存技術)を行い、結果として、PnとPkがリンクされる。但し、n≠k (n++)
474
係り受け抽出ルール
470
係り受けの抽出が終了(n=i)
Yes
No
480
終 了
472
ルールに基づきPnに対してリンクされた２項関係を抽出する (n++)

【図５】


相対頻度
計算装置
530
入力（命令）
装置
570
命令
解析装置
520
表示部
510
キーワード
検索装置
540
カテゴリー別
キーワード頻度
計算装置
550
ラベル・
概念付き
データベース
560

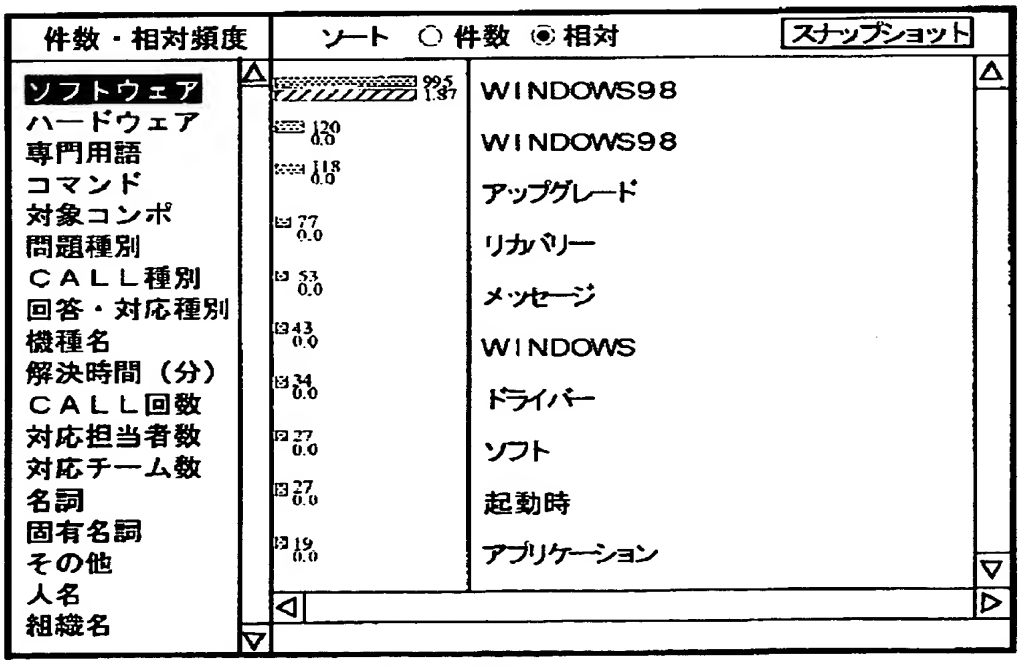
【図６】

概念Ａ

概念Ｂ	A1	A2	A3	A4	A5	A6	A7
B1	1720 (89.6%)	56 (2.9%)	118 (6.1%)	7 (0.3%)	8 (0.4%)	7 (0.3%)	1 (0.0%)
B2	1255 (90.5%)	21 (1.5%)	100 (7.2%)	4 (0.2%)	2 (0.1%)	2 (0.1%)	2 (0.1%)
B3	1072 (95.0%)	28 (2.4%)	18 (1.5%)	5 (0.4%)	0 (0.0%)	3 (0.2%)	1 (0.0%)
B4	501 (90.1%)	9 (1.6%)	40 (7.1%)	1 (0.1%)	3 (0.5%)	2 (0.3%)	0 (0.0%)
B5	649 (93.2%)	26 (3.7%)	17 (2.4%)	1 (0.1%)	0 (0.0%)	3 (0.4%)	0 (0.0%)
B6	682 (95.5%)	13 (1.8%)	15 (2.1%)	3 (0.4%)	0 (0.0%)	1 (0.1%)	0 (0.0%)
B7	676 (96.9%)	11 (1.5%)	5 (0.7%)	3 (0.4%)	1 (0.1%)	1 (0.1%)	0 (0.0%)
B8	673 (90.7%)	34 (4.5%)	26 (3.5%)	4 (0.5%)	0 (0.0%)	5 (0.6%)	0 (0.0%)
B9	528 (74.3%)	134 (18.8%)	10 (1.4%)	14 (1.9%)	2 (0.2%)	22 (3.0%)	0 (0.0%)
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

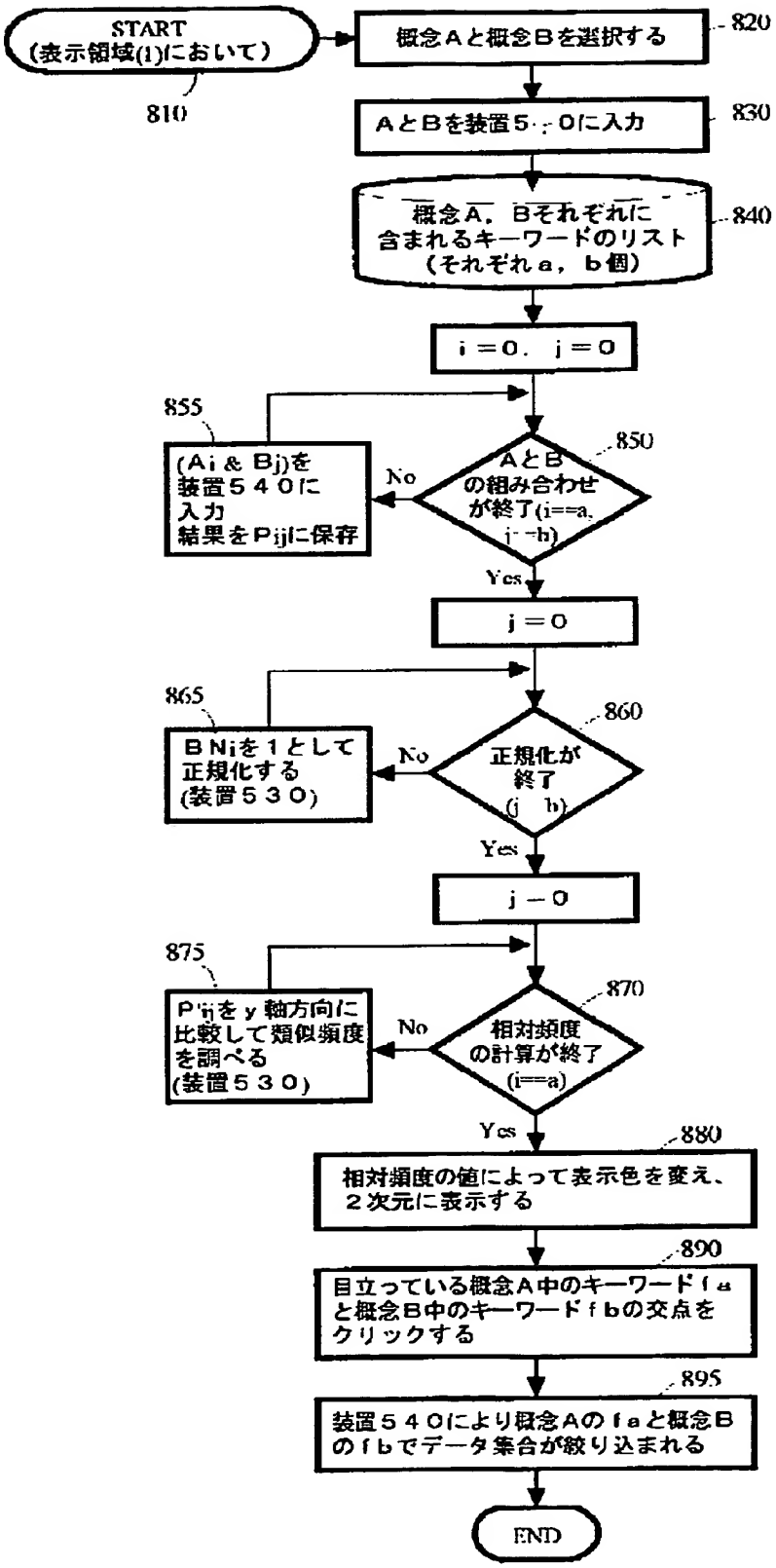
表示領域（１）

【図７】

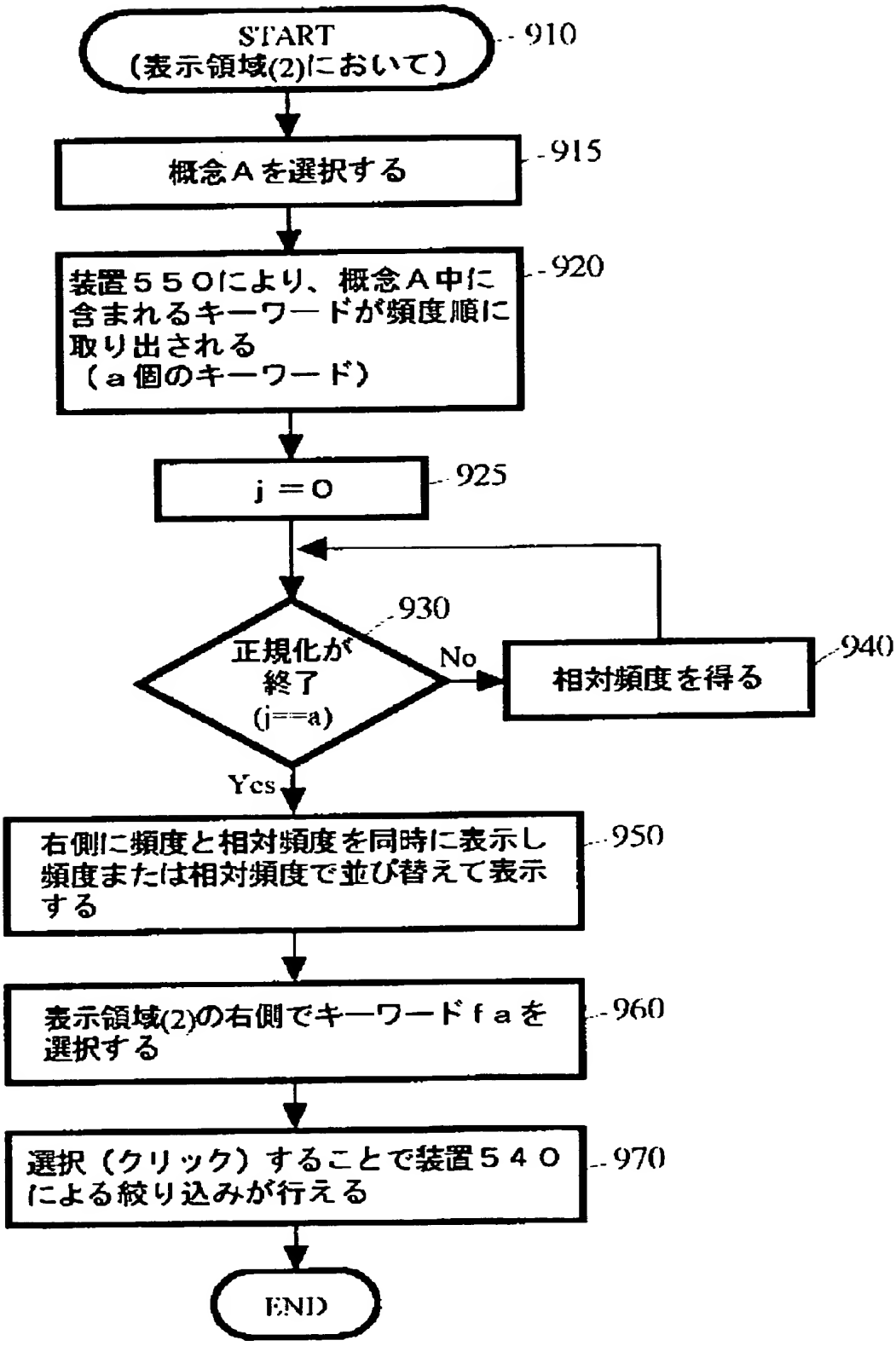


表示領域（２）

【図８】


START
(表示領域(1)において)
810
概念Ａと概念Ｂを選択する
820
ＡとＢを装置５１０に入力
830
概念Ａ，Ｂそれぞれに
含まれるキーワードのリスト
（それぞれａ，ｂ個）
840
ｉ＝０，ｊ＝０
855
(Ａｉ＆Ｂｊ)を
装置５４０に
入力
結果をＰｉｊに保存
No
850
ＡとＢ
の組み合わせ
が終了(i==a,
j==b)
Yes
ｊ＝０
865
ＢＮｉを１として
正規化する
(装置５３０)
No
860
正規化が
終了
(j==b)
Yes
ｊ＝０
875
Ｐｉｊをｙ軸方向に
比較して類似頻度
を調べる
(装置５３０)
No
870
相対頻度
の計算が終了
(i==a)
Yes
880
相対頻度の値によって表示色を変え、
２次元に表示する
890
目立っている概念Ａ中のキーワードｆａ
と概念Ｂ中のキーワードｆｂの交点を
クリックする
895
装置５４０により概念Ａのｆａと概念Ｂ
のｆｂでデータ集合が絞り込まれる
END

【図９】

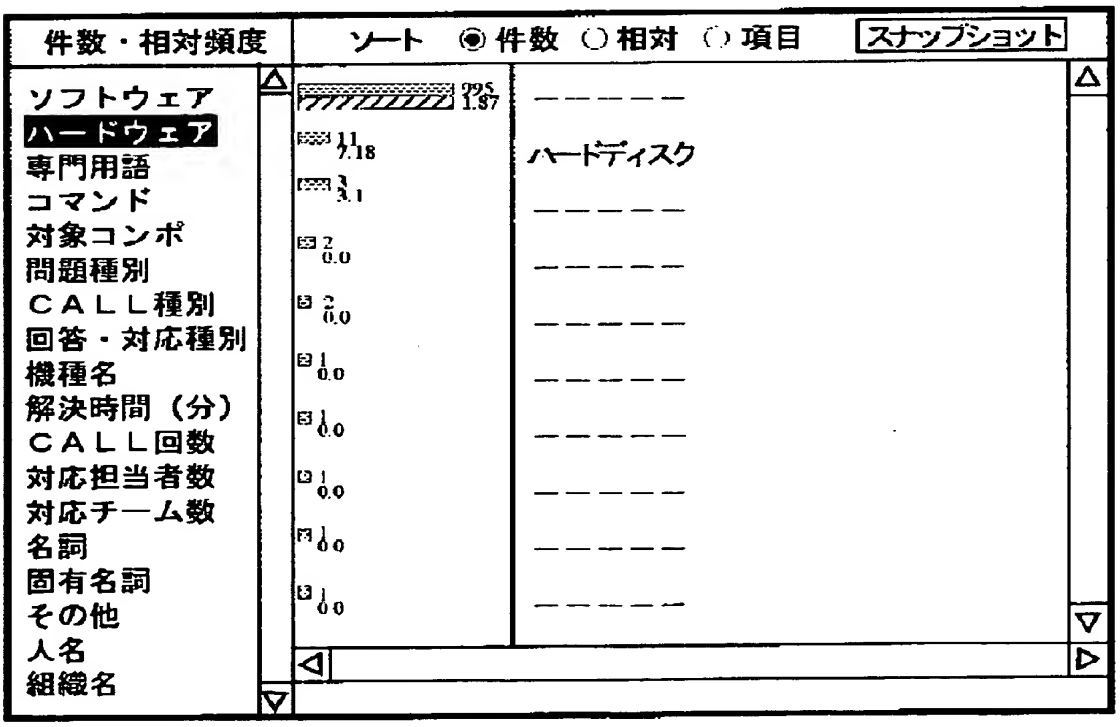

START
(表示領域(2)において)
910
概念Ａを選択する
915
装置５５０により、概念Ａ中に含まれるキーワードが頻度順に取り出される
（ａ個のキーワード）
920
ｊ＝０
925
930
正規化が終了
(j==a)
No
相対頻度を得る
940
Yes
右側に頻度と相対頻度を同時に表示し頻度または相対頻度で並び替えて表示する
950
表示領域(2)の右側でキーワードｆａを選択する
960
選択（クリック）することで装置５４０による絞り込みが行える
970
END

【図１０】

						遅い	
	19 (2.58%)	6 (0.82%)	9 (1.22%)	6 (0.82%)	2 (0.27%)	2 (0.27%)	3 (0.41%)
	18 (2.02%)	7 (0.79%)	18 (2.02%)	4 (0.45%)	4 (0.45%)	3 (0.34%)	3 (0.34%)
	21 (2.75%)	14 (1.83%)	15 (1.97%)	12 (1.57%)	12 (1.57%)	14 (1.83%)	7 (0.92%)
	57 (5.76%)	12 (1.21%)	11 (1.11%)	8 (0.81%)	3 (0.31%)	1 (0.10%)	5 (0.51%)
	55 (4.8%)	20 (1.75%)	7 (0.61%)	11 (0.96%)	2 (0.17%)	1 (0.95%)	5 (0.44%)
製品Ａ	20 (2.58%)	15 (1.94%)	14 (1.81%)	8 (1.03%)	6 (0.78%)	20 (2.58%)	6 (0.78%)
	15 (2.11%)	18 (2.53%)	12 (1.69%)	9 (1.27%)	1 (0.14%)	5 (0.70%)	3 (0.42%)
	17 (2.29%)	10 (1.35%)	33 (4.45%)	6 (0.81%)	1 (0.13%)	2 (0.27%)	2 (0.27%)
	15 (2.76%)	7 (1.29%)	8 (1.47%)	5 (0.92%)	2 (0.37%)	0 (0.00%)	3 (0.55%)
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

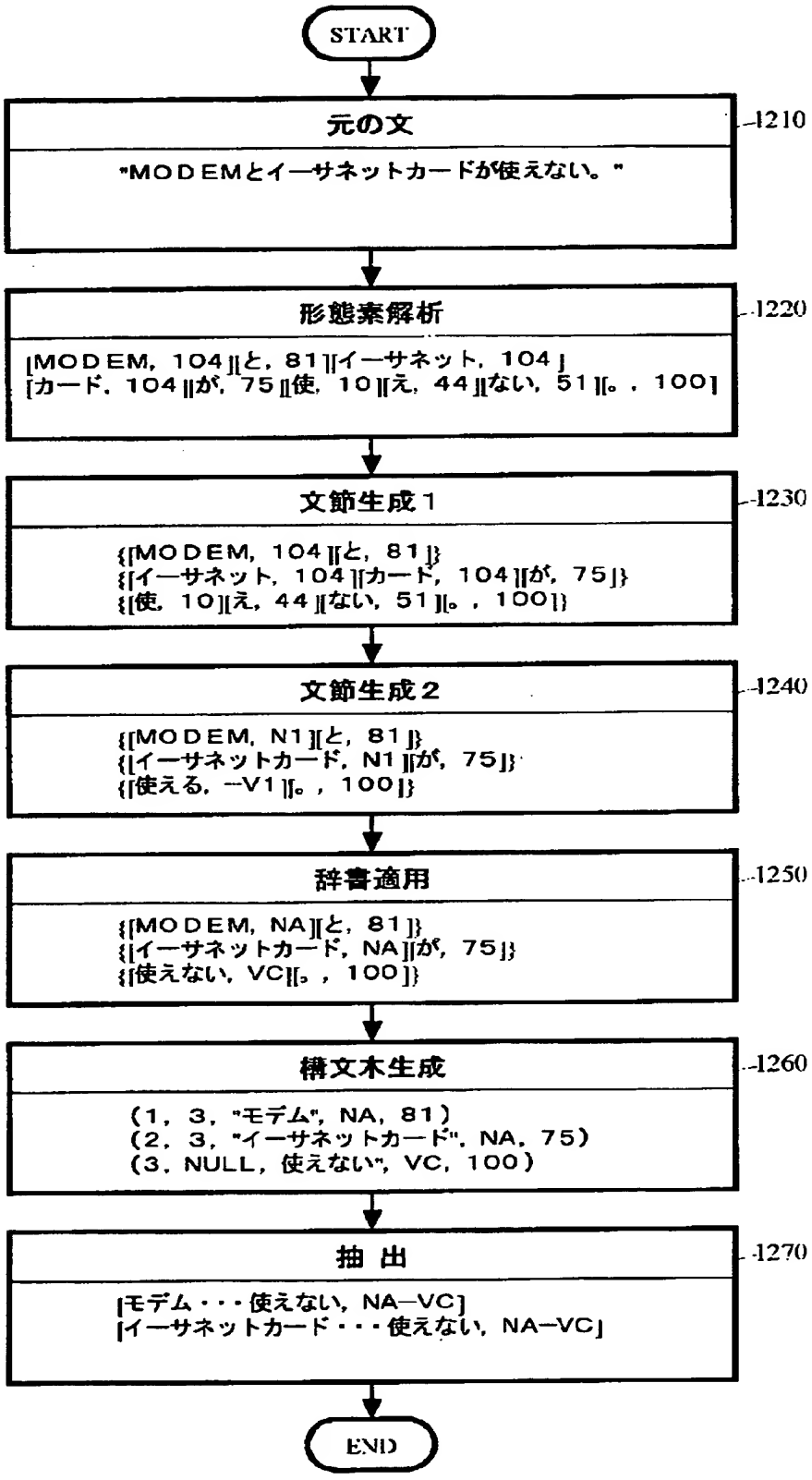
表示領域（１）

【図１１】

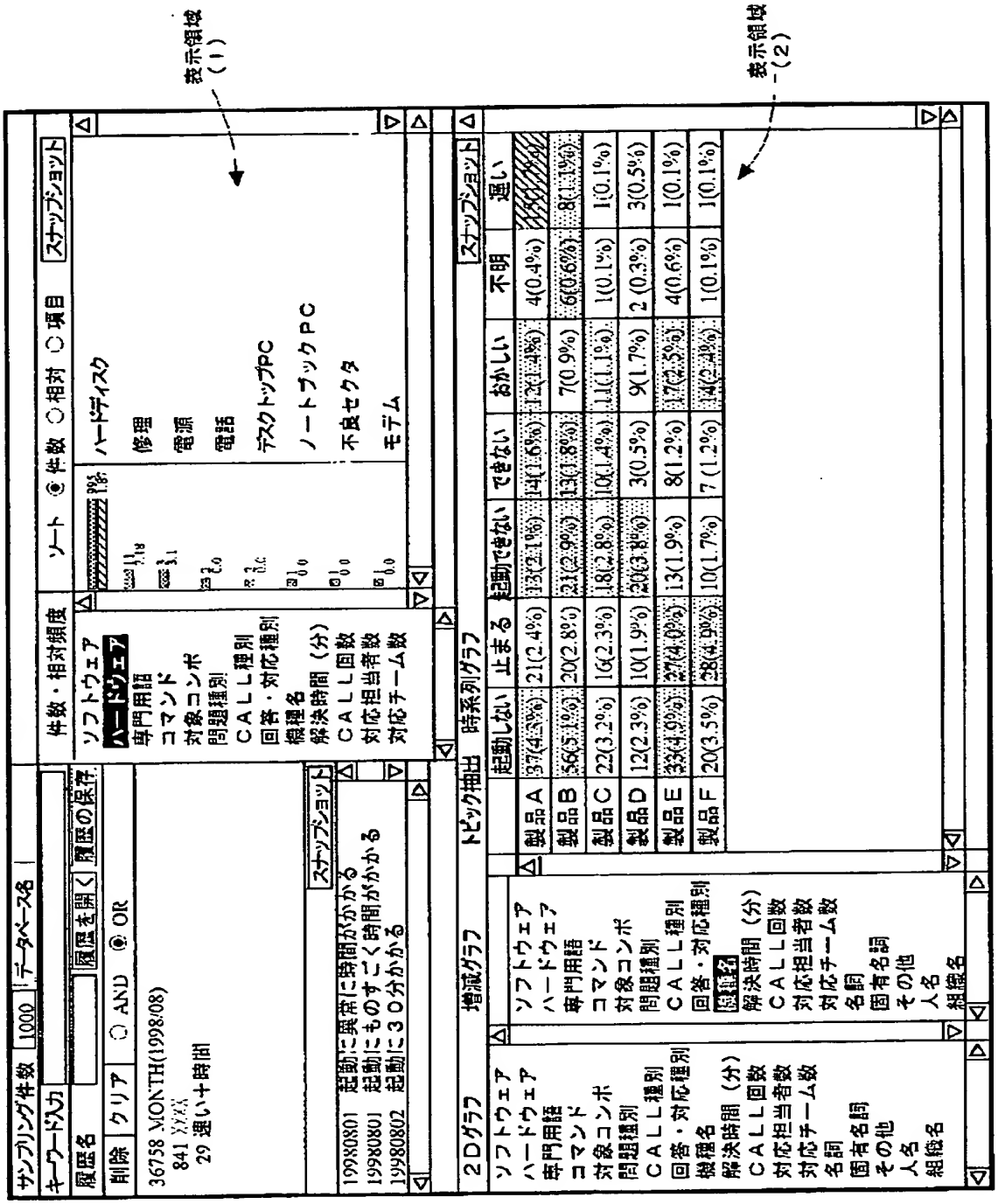


表示領域（２）

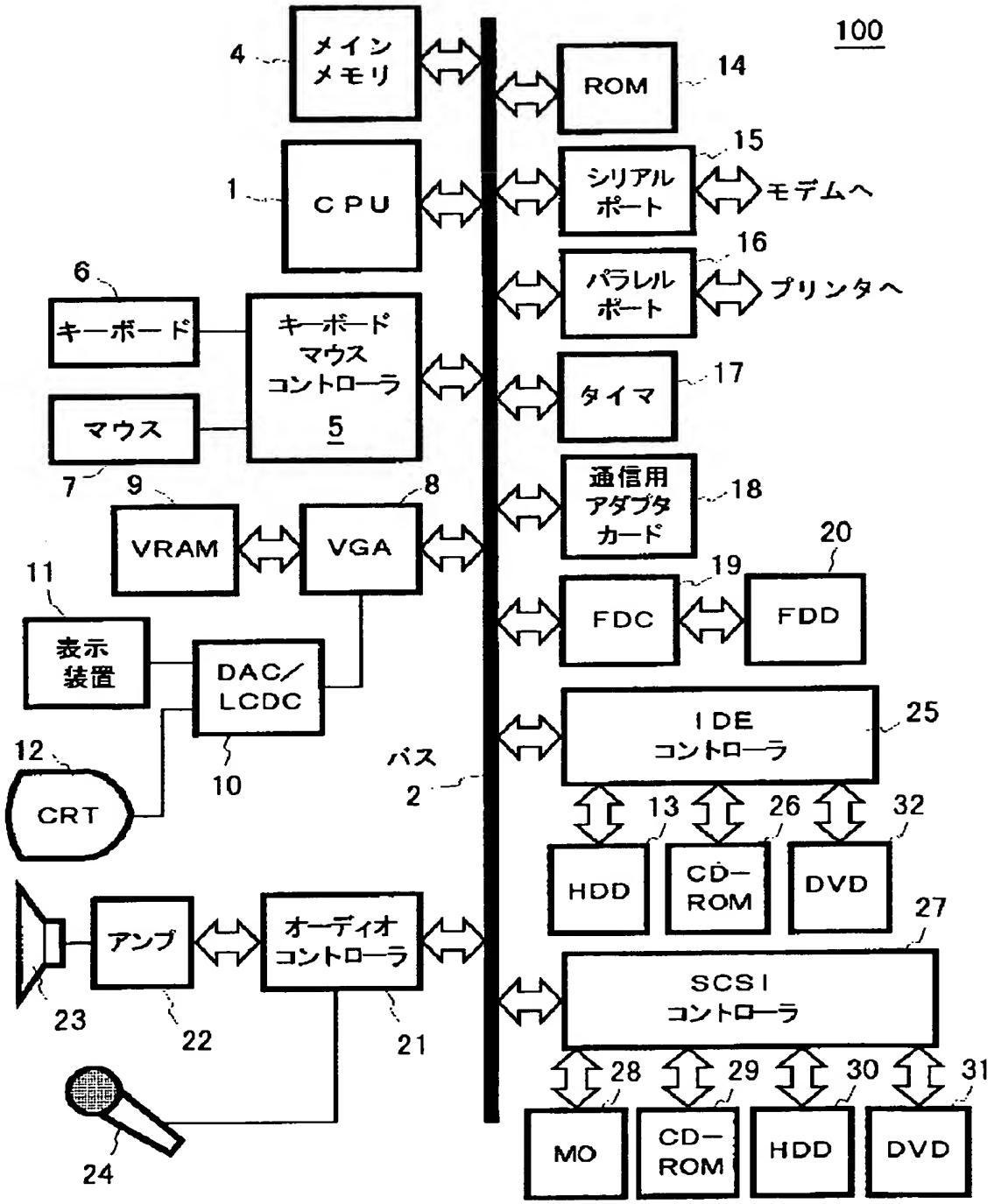
【図１２】


START
元の文
"MODEMとイーサネットカードが使えない。"
1210
形態素解析
[MODEM, 104][と, 81][イーサネット, 104]
[カード, 104][が, 75][使, 10][え, 44][ない, 51][。, 100]
1220
文節生成１
{[MODEM, 104][と, 81]}
{[イーサネット, 104][カード, 104][が, 75]}
{[使, 10][え, 44][ない, 51][。, 100]}
1230
文節生成２
{[MODEM, N1][と, 81]}
{[イーサネットカード, N1][が, 75]}
{[使える, -V1][。, 100]}
1240
辞書適用
{[MODEM, NA][と, 81]}
{[イーサネットカード, NA][が, 75]}
{[使えない, VC][。, 100]}
1250
構文木生成
(1, 3, "モデム", NA, 81)
(2, 3, "イーサネットカード", NA, 75)
(3, NULL, 使えない", VC, 100)
1260
抽 出
[モデム・・・使えない, NA-VC]
[イーサネットカード・・・使えない, NA-VC]
1270
END

【図１３】


表示領域（１）
表示領域（２）
サンプリング件数 1000 データベース名
キーワード入力
履歴名 履歴を開く 履歴の保存
削除 クリア AND OR
36758 MONTH(1998/08)
841 XXXX
29 遅い＋時間
スナップショット
1998080101 起動に異常に時間がかかる
1998080801 起動にものすごく時間がかかる
1998080202 起動に３０分かかる
スナップショット
2Dグラフ 増減グラフ トピック抽出 時系列グラフ
ソフトウェア
ハードウェア
専門用語
コマンド
対象コンポ
問題種別
ＣＡＬＬ種別
回答・対応種別
機種名
解決時間（分）
ＣＡＬＬ回数
対応担当者数
対応チーム数
名詞
固有名詞
その他
人名
組織名
件数・相対頻度 ソート 件数 相対 項目
ソフトウェア
ハードウェア
専門用語
コマンド
対象コンポ
問題種別
ＣＡＬＬ種別
回答・対応種別
機種名
解決時間（分）
ＣＡＬＬ回数
対応担当者数
対応チーム数
ハードディスク
修理
電源
電話
デスクトップPC
ノートブックPC
不良セクタ
モデム
スナップショット
起動しない 止まる 起動できない できない おかしい 不明 遅い
製品Ａ 37(4.3%) 21(2.4%) 13(2.1%) 14(1.6%) 12(1.4%) 4(0.4%)
製品Ｂ 56(6.1%) 20(2.8%) 21(2.9%) 13(1.8%) 7(0.9%) 6(0.6%) 8(1.1%)
製品Ｃ 22(3.2%) 16(2.3%) 18(2.8%) 10(1.4%) 11(1.1%) 1(0.1%) 1(0.1%)
製品Ｄ 12(2.3%) 10(1.9%) 20(3.8%) 3(0.5%) 9(1.7%) 2(0.3%) 3(0.5%)
製品Ｅ 33(4.9%) 27(4.0%) 13(1.9%) 8(1.2%) 17(2.3%) 4(0.6%) 1(0.1%)
製品Ｆ 20(3.5%) 28(4.9%) 10(1.7%) 7(1.2%) 14(2.4%) 1(0.1%) 1(0.1%)

【図１４】


100
メイン
メモリ
4
ROM
14
CPU
1
シリアル
ポート
15
モデムへ
パラレル
ポート
16
プリンタへ
6
キーボード
キーボード
マウス
コントローラ
5
タイマ
17
マウス
7
9
8
通信用
アダプタ
カード
18
VRAM
VGA
20
19
FDC
FDD
11
表示
装置
DAC/
LCDC
IDE
コントローラ
25
バス
2
12
CRT
10
13
26
32
HDD
CD-
ROM
DVD
27
アンプ
オーディオ
コントローラ
SCSI
コントローラ
23
22
21
28
29
30
31
MO
CD-
ROM
HDD
DVD
24

【書類名】　要約書

【要約】

【課題】　　大量データから有効な知識を獲得する方法およびシステムを提供することである。

【解決手段】

本発明は文書を含む大量のデータから特異な特徴を有する概念を抽出することにより、有効な知識を獲得する方法およびシステムを提供する。本発明は概念抽出装置と特徴的概念抽出装置と有する。前記概念抽出は文書データを含むデータからカテゴリ別の概念を抽出する。前記特徴的概念抽出装置は前記抽出した概念の中から特徴的な概念を抽出する装置であって、前記カテゴリ別の概念において、同一カテゴリに属する概念のうち、対応する別のカテゴリに属する概念の中で占める割合が既定値を超えている概念を抽出する。

【選択図】　図１

出証特平１１－３０８３３３６

認定・付加情報

特許出願の番号	平成１１年　特許願　第１９２６１８号
受付番号	５９９００６５０４４９
書類名	特許願
担当官	第七担当上席　　００９６
作成日	平成１１年　７月　９日

＜認定情報・付加情報＞

【提出日】　平成11年 7月 7日

次頁無

出　願　人　履　歴　情　報

識別番号　　〔３９０００９５３１〕

１．変更年月日　　１９９０年１０月２４日
［変更理由］　新規登録
住　所　　アメリカ合衆国１０５０４、ニューヨーク州　アーモンク　（番地なし）
氏　名　　インターナショナル・ビジネス・マシーンズ・コーポレイション